ENVRI Common Operations of Environmental Research Infrastructures

# ENVRI

## Services for the Environmental Community

# Analysis of Common Requirements For ENVRI Research Infrastructures

| | |
|---|---|
| Document identifier: | D3.3 Analysis of Common Requirements for ENVRI Research Infrastructures |
| Date: | **30/04/2013** |
| Activity: | **WP3** |
| Lead Partner: | **CU** |
| Document Status: | **FINAL** |
| Dissemination Level: | **PUBLIC** |
| Document Link: | <link to the website> |

## ABSTRACT

The objectives of ENVRI WP3 task T3.2 is to examine the design of the 6 ESFRI environmental Research Infrastructures (RIs), (ICOS, EURO-Argo, EISCAT-3D, LifeWatch, EPOS, and EMSO,) in order to identify common computational characteristics of them, and to develop an understanding of the specific requirement through observations.

Throughout the study, a standard model, the Open Distributed Processing (ODP) is chosen to use to interpret the design of the research infrastructures, and place their requirements into the ODP framework for analysing. The document reports the initial results from this study. Briefly, from the aspect of the ODP Engineering Viewpoint, the architectural characteristics of the RIs have been examined, and 5 common sub-systems have been identified: sub-systems of **data acquisition**, **curation**, **access**, **processing** and **community support**. Secondly, from the aspect of the ODP Computational Viewpoint, we looked at each of the 6 RIs in details and identified the common functions and embedded computations they provided. Matrices has been used for comparison. Definitions of functionalities have been provided. Finally, from the aspect of the ODP enterprise viewpoint, we have identified 4 common communities, and derived the community roles.

The contribution of this work to the environmental science research infrastructures is threefold:

- It investigates 6 ESFRI RIs, which is a collection of representative research infrastructures for environmental sciences, and provides a projection of Europe-wide requirements they have; identifying in particular, those they have in common;
- It experiments with ODP as an approach in requirement analysis, to serve as a common language for interpretation and discussion to ensure unifying understanding;
- The results from this study can be used as an input to a design or implementation model. Common services can be provided in the light of the common analysis, which can be widely applicable to various environmental research infrastructures and beyond.

ENVRI Common Operations of Environmental Research Infrastructures

## 1. COPYRIGHT NOTICE

## 2. DELIVERY SLIP

| | Name | Partner/Activity | Date |
|---|---|---|---|
| **From** | | | |
| **Reviewed by** | **Moderator:**<br>**Reviewers:** | | |
| **Approved by** | | | |

## 3. DOCUMENT LOG

| Issue | Date | Comment | Author/Partner |
|---|---|---|---|
| 1.0 | 17/02/13 | First draft for internal review | Yin Chen (CU)<br>Alex Hardisty (CU)<br>Alun Preece (CU)<br>Paul Martine (UEDIN)<br>Malcolm Atkinson (UEDIN)<br>Herbert Schentz (EAA)<br>Barbara Magagna(EAA)<br>Zhiming Zhao(UoV) |
| 2.0 | 30/04/13 | Internally reviewed version to be approved by project management and submitted to the Commission. | Yin Chen (CU) |
| 3.0 | | | |
| 4.0 | | | |
| 5.0 | | | |
| 6.0 | | | |

## 4. APPLICATION AREA

This document is a formal deliverable for the European Commission, applicable to all members of the ENVRI project, beneficiaries and Joint Research Unit members, as well as its collaborating projects.

## 5. DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the authors.

## 6. TERMINOLOGY

A complete project glossary is provided at the following page: http://www.ENVRI.eu/glossary.

## 7. PROJECT SUMMARY

Frontier environmental research increasingly depends on a wide range of data and advanced capabilities to process and analyse them. The ENVRI project, “Common Operations of Environmental Research infrastructures”, is a collaboration in the ESFRI Environment Cluster, with support from ICT experts, to develop common e-science components and services for their facilities. The results will speed up the construction of these infrastructures and will allow scientists to use the data and software from each facility to enable multi-disciplinary science.

The target is on developing common capabilities including software and services of the environmental e-infrastructure communities. While the ENVRI infrastructures are very diverse, they face common challenges including data capture from distributed sensors, metadata standardisation, management of high volume data, workflow execution and data visualisation. The common standards, deployable services and tools developed will be adopted by each infrastructure as it progresses through its construction phase.

Two use cases, led by the most mature infrastructures, will focus the development work on separate requirements and solutions for data pre-processing of primary data and post-processing toward publishing.

The project will be based on a common reference model created by capturing the semantic resources of each ESFRI-ENV infrastructure. This model and the development driven by the test-bed deployments result in ready-to-use systems which can be integrated into the environmental research infrastructures.

The project puts emphasis on synergy between advanced developments, not only among the infrastructure facilities, but also with ICT providers and related e-science initiatives. These links will facilitate system deployment and the training of future researchers, and ensure that the inter-disciplinary capabilities established here remain sustainable beyond the lifetime of the project.

## 8. EXECUTIVE SUMMARY

This report presents the initial finding from the study T3.2, analysis of the requirements of data processing. The Open Distributed Processing (ODP) is used as the framework for the analysis. From the aspect of the ODP *engineering viewpoint*, the physical structuring mechanism for the 6 ENVRI research infrastructures are analysed and 5 common sub-systems are identified, *data acquisition*, *data curation*, *data access*, *data processing*, and *user community support*. Secondly, from the aspect of the *computational viewpoint*, a set of operations and embedded computations commonly provided by the infrastructures are identified. Finally, from the aspect of the ODP *enterprise viewpoint*, 4 common *communities* are identified: *data acquisition*, *data management*, *data service provision*, and *data user*. The roles, behaviours, and policies for each community are described.

# TABLE OF CONTENTS

# 1 INTRODUCTION

The objective of this study is to analyse and develop an understanding of the specific requirements of each ESFRI Environmental (ENV) Research Infrastructure (RI) with respect to common short-term priorities. The ENVRI background papers and T3.1 Assessment of the State of the Art provide useful surveys and evaluations of ENV RIs; common requirements emerge. This study intends to make a further step towards a common model for ENVRI. The study will not produce a common model -- however, it will serve as an input to such a model.

In this study, we use a standard approach, Open Distributed Processing (ODP), to interpret the design of 6 representative environmental research infrastructures (ICOS[1], EPOS[2], EMSO[3], EISCAT-3D[4], LifeWatch[5], and Euro-Argo[6]), and place their requirements into the ODP framework for further analysis. ODP is an ISO/IEC standard [1-4], which provides an overall conceptual framework for building distributed system. It defines five specific viewpoints which are abstractions that yield specifications of the whole system related to particular sets of concerns. The five viewpoints are [5];

- **The Enterprise Viewpoint**, which concerns the organisational situation in which design activity is to take place;
- **The Information Viewpoint**, which concerns the modelling of the shared information manipulated within the infrastructure of interest;
- **The Computational Viewpoint**, which concerns the development of the high-level design of the processes and applications supporting the infrastructure;
- **The Engineering Viewpoint**, which tackles the problems of diversity in infrastructure provision; it gives the prescriptions for supporting the necessary abstract computational interactions in a range of different situations;
- **The Technology Viewpoint**, which concerns with managing real-world constraints, such as restrictions on the facilities available to implement the system, to the existing application platforms on which the applications must run.

The added value of using ODP to analyse the requirements for ENVRI is threefold:

- To use a standardised language to interpret the design of research infrastructures which helps to unify understanding;
- To provide a gentler pathway to link real-world systems to a ODP model world;
- Since the ODP framework is created to help designers deliver a practical architecture which leads to concrete implementations, using ODP concepts for requirements analysis can help us to drill down to details and identify essential problems.

The analysis presented in this reports are based on partial knowledge of a snapshot of current state of ENV RIs. This is because the documentation of RIs is often incomplete and inconsistent, and the

[1] ICOS: http://www.icos-infrastructure.eu/

[2] EPOS: http://www.epos-eu.org/

[3] EMSO: http://www.emso-eu.org/management/

[4] EISCAT-3D: http://www.eiscat3d.se/

[5] LifeWatch: http://www.lifewatch.eu/

[6] Euro-Argo: http://www.euro-argo.eu/

designs evolve over time and subject to change. For example, the EISCAT-3D design study finished in 2009 and submitted the final design as Deliverable 11.1 to the EU commission. Its succeeding project, the EISCAT-3D Preparatory Phase (EISCAT-3D PP), started in 2010, examines the feasibility of the design and prepares for implementation starting in 2014. During the EISCAT-3D PP, many parts of the design are likely to be re-evaluated and re-designed, e.g., due to infeasibility for implementation. This is a common issue for most if not all other RIs. The investigation of ENVRI should be based on the existing knowledge of the design provided by RIs, meanwhile, keep up to date with the development of any new activities.

There is one key issue of how to denote in the ODP context "known unknowns". We should tolerate schemas/descriptions with many of these at first, and progressively push the unknowns towards detail or boundaries later.

The rest of the report presents initial findings from the study. The analysis covers 3 ODP viewpoints, *Enterprise*, *Computational*, and *Engineering*. Because most of the ENV RIs provide documentation which describes the architectural features of their infrastructure, it is straightforward to start with the *Engineering Viewpoint*, where we identify the common physical structuring mechanism for the system infrastructures. Secondly, we identify common functions and embedded computations provided by the ENV RIs. This, in essence, is to analyse the RIs from the aspect of the ODP *Computational Viewpoint*. Finally, we look at the real-world systems from the aspect of the ODP *Enterprise Viewpoint*, and identify the common *communities*, and their *roles*. Matrices are used to visualise the results of comparison, where columns are the names of ENV RIs and rows are ODP elements.

# 2 COMMON ARCHITECTUREAL CHARACTERISTICS

In ODP, the purpose of the *Engineering Viewpoint* is to identify and specify the structuring mechanisms for distributed interactions and the functional elements. It concerns the architectural features of an infrastructure.

The structures of the studied RIs can be divided into **sub-systems** based on functions and locations of computational elements. For the purposes of this document, each *sub-system* is defined as a set of capabilities that collectively are defined by a set of **interfaces** with corresponding operations that can be invoked by other *sub-systems*. An interface in ODP is an abstraction of the behaviour of an object that consists of a subset of the interactions of that object together with a set of constraints on when they may occur. *Sub-systems* are disjoint from each other.

Five common *sub-systems* are identified: **data acquisition**, **data curation**, **data access**, **data processing**, and **community support**. The order of these *sub-systems* is irrelevant.

The **data acquisition sub-system** collects raw data from sensor arrays, various instruments, or human observers, and brings the measures (data streams) into the system. Note, ENVRI is concerned with the computational aspects of an infrastructure, thus, by definition, the *data acquisition sub-system* starts from the point of sensor signals being converted into digital values and received by the system. There are many related activities including, defining data acquisition protocols, design and deployment of the sensor instruments, and configuration and calibration devices, which are crucial tasks for data acquisition nevertheless beyond the scope of the ENVRI investigation. The *data acquisition sub-system* is typically operated at observatories or stations. Data in the *acquisition sub-system* are normally non-reproducible, the so-called raw data or primary data. Consistent time-stamps are assigned to each data object. There are the cases that the raw data may be generated by a simulation model, in which situation, the raw data may be reproducible, in terms of being regenerated. The (real-time) data streams sometimes are temporarily stored (e.g., in computer clusters), then, sampled, filtered or processed (e.g., based on applied quality control criteria). Control software is often provided to allow the execution and monitoring of data flows. The data collected at the *data acquisition sub-system* are transmitted to the *data curation sub-system*, to be maintained and archived there.

The **data curation sub-system** facilitates quality control and preservation of scientific data. It is typically operated at a data centre. Data handled at the *curation sub-system* are often reproducible in term of being able to be re-processed. Operations such as data quality verification, data identification, annotation, cataloguing, and long-term preservation are often provided. Various data products are generated and provided for users which need to be accessed through data *access sub-system*. There is usually an emphasis on non-functional requirements for a *data curation sub-system* including the need for satisfying performance criteria in availability, reliability, utility, throughput, responsiveness, security and scalability.

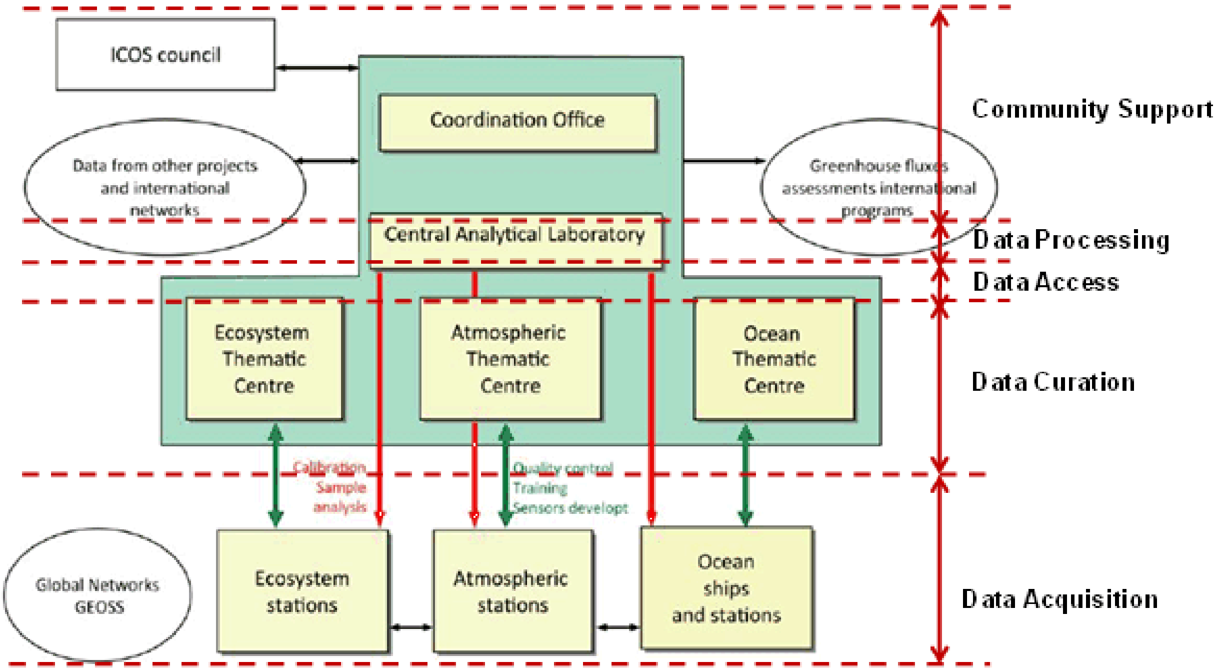
The **data access sub-system** enables discovery and retrieval of data housed in data resources managed by a *data curation sub-system*. *Data access sub-systems* often provide facilities such as data portals, as well as services to present or deliver the data products. Search facilities including both query-based and navigation-based searching tools are provided which allow users or services to discover interesting data products. Discoveries based on metadata or semantic linkages are most common. Data handled at the *access sub-system* can be either structurally and semantically homogeneous or heterogeneous. When supporting heterogeneous data, different types of data (often pulled from a variety of distributed data resources) may be converted into uniform representations with uniform semantics which can be resolved by a data discovery and access service. Services allowing harvesting of metadata and/or data, as well as services enhancing the performance by compression and packaging methods and encoding services for secure data transfer are often part of the *data access sub-system*. Data access can be open or controlled (e.g., enforced by authentication and authorisation policies). It is notable that a *data access sub-system* usually does not provide "write" operations for end users, although such operations may be provided for an administrator of a data resource.

The **data processing sub-system** aggregates the data from various resources and provides computational capabilities and capacities for conducting data analysis and scientific experiments. Data handled by the *data processing sub-system* are typically derived and recombined via the *data access sub-system*. A *data processing sub-system* normally offers operations for statistical and/or mining functions for analysis, facilities for conducting scientific experiments, modelling/simulation, and scientific visualisation. Performance requirements for processing scientific data tend to be concerned more about scalability issue, which may also be necessary to address at the infrastructure level -- for example, to make use of the Grid or Cloud technology. In this case, functionalities to interact with the physical infrastructure should be provided.

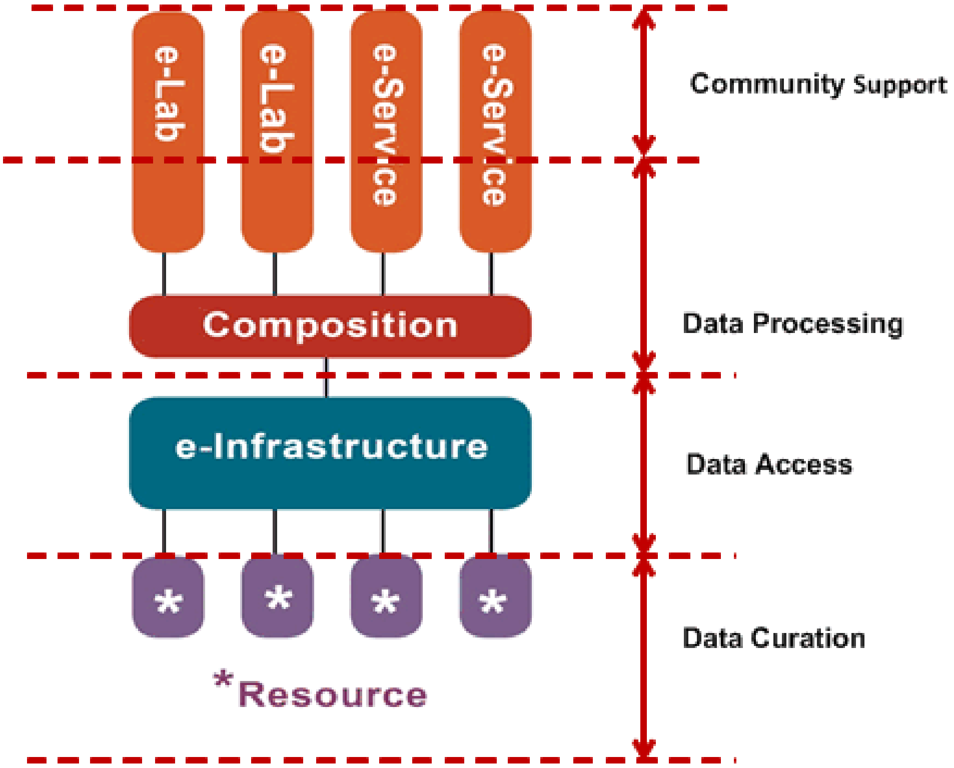
Finally, the **community support sub-system** manages, controls and tracks users' activities and supports users to conduct their roles in communities. Data handled by a *community support sub-system* typically are user generated data, control and communications. A *community support sub-system* normally supports for interactive visualisations, Authentication, Authorisation and Accounting (AAA), as well as for managing virtual organisations. The *community support* is orthogonal to and cross-cutting the other 4 *sub-systems*.

There may be other ways to group the functional elements. Above provides one possible solution. The main purpose for the classification is to identify the common structural characteristics of the environmental research infrastructures. As shown in Figure 2.1 below, the five *sub-systems* map well to the architectures of the RIs studied.

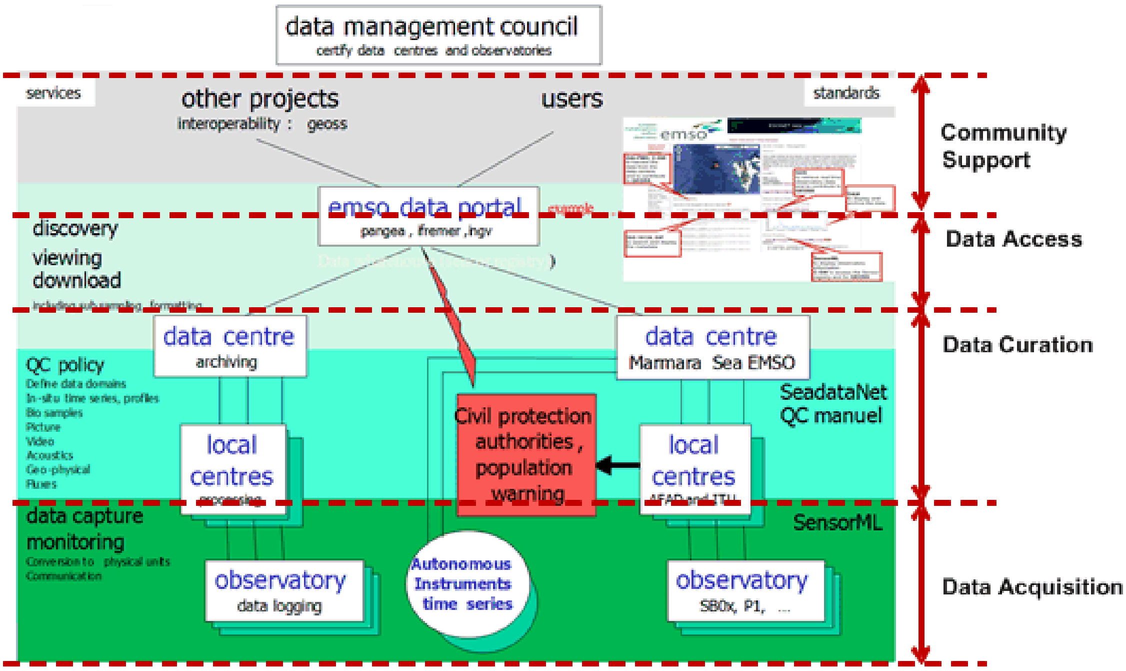
As shown in Table 2.1, different RIs emphasise the design and implementation of different *sub-systems*. By the time of writing this report, RIs such as ICOS, EISCAT-3D, Euro-Argo, and EMSO mainly focus on data *acquisition*, *curation* and *access*. They are typical **large-scale observatory systems**. Some others RIs, such as EPOS and LifeWatch, are built on existing systems having limited control over data resources, and focus more on data *access* and *processing*. They are **comprehensive integration infrastructures** for domain data and computations. It worthy of mentioning that generic computational RIs, such as EUDAT and EGI, are **general purpose large-scale infrastructures** for data management and processing; EUDAT tends to focus more on the functionalities related to the data *curation sub-system*, and EGI tends to focus more on the functionalities related to the data *processing sub-system*. Both EUDAT and EGI provide generic operations and services which can be used in various domains of research within either infrastructure.



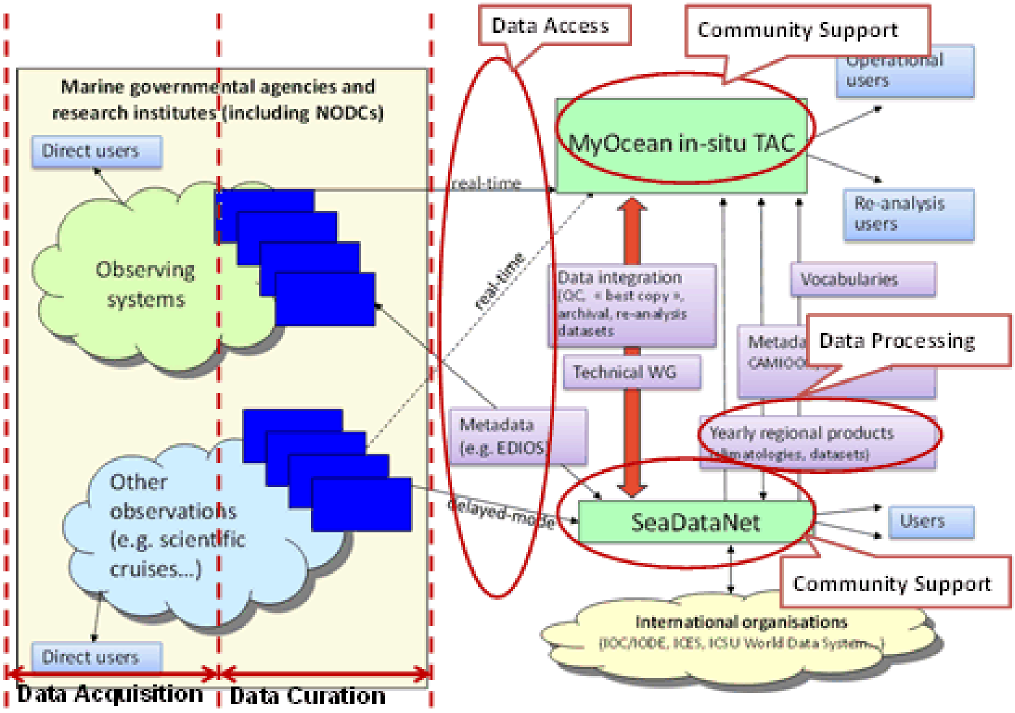
(A) ICOS Architecture



(B) LifeWatch Architecture



(C) EMSO Architecture



(D) Euro-Argo Architecture

Figure 2.1: Common Sub-Systems

The *community support sub-system* is commonly requested by users of new RIs. For example, in EISCAT-3D, there is a need to allow users to control remote radar systems to collect their own data; Both EURO-Argo and ICOS request the expertise of users to verify the quality of data; EMSO (PANGAEA system) facilitates user to submit metadata; whilst EPOS and LifeWatch plan to allow users to share their data and experiments workflows. However, currently we observe only few RIs actively designing or implementing this *sub-system*. It is likely because of resource limitations. EGI offers more experience of supporting its user community. Operations, such as Virtual Organizations Management Service (VOMS) and accounting are provided through EGI infrastructure services. However, there are no generic tools or software available to support many newly emerged requirements, such as community coordination, collaboration, policy making, and user collaborative work and publication of results. We expect the RIs with greater maturity and sufficient resources will lead the way to explore this area in the near future. In the rest of the study, we will not examine the requirements for this *sub-system* intensively.

Table 2.1: The Correlations Between the Design and Implementation Emphasis of ENV RIs and the Five Common Sub-systems

| **Sub-system** | **EISCAT-3D** | **Euro-Argo** | **ICOS** | **EMSO** | **EPOS** | **LifeWatch** | **EUDAT** | **EGI** |
|---|---|---|---|---|---|---|---|---|
| **Acquisition** | Yes | Yes | Yes | Yes | | | | |
| **Curation** | Yes | Yes | Yes | Yes | | | Yes | |
| **Access** | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| **Processing** | | | | | Yes | Yes | | Yes |
| **Community Support** | | | | | | Partial | Partial | Partial |

For the same *sub-system*, different RIs provides different facilities. In the following, we examine each RI, and identify the common computational functions in each *sub-system*.

# 3 COMMON FUNCTIONS AND OPERATIONS

The ODP *Computational Viewpoint* focuses on the functionality of an infrastructure, and the service it offers.

Dividing the structures of RIs into *sub-systems* helps to break down the complexity in analysis. Within each *sub-system*, we use a data-oriented approach, which follows the life-cycle of data -- e.g., creation, transmission, transformation, modification, processing, and visualisation -- to identify key functions and embedded computations.

The analysis is based on the materials or information from the following sources:

- ENVRI background papers
- ENVRI Deliverable D3.2 Assessment of the State of the Art
- Published papers, deliverables, white papers and websites of investigated RIs, and
- Interviews with RIs

## 3.1 Analysis of EISCAT-3D

The objective of EISCAT-3D [4] is to design and construct a new-generation incoherent-scatter research radar which provides a long-term upper atmospheric science capability for studies of the atmosphere and near-Earth space.

The system design of EISCAT-3D explored many different areas, including the construction of antennas, arrays, the signal processing, the network and the data distribution system. The investigation of ENVRI scopes into data acquisition, processing and archiving aspects of EISCAT-3D.

The design of EISCAT-3D data archiving and distribution can be summarised as follows: there is a two-stage system for handling data. The beam-formed sample-level data, together with data from the interferometry system, and some high-volume data from supporting instruments, are streamed to a large ring buffer designed to hold up to a few days, after which these low-level data will be over-written. The ring buffer allows the low-level data to be stored for long enough to allow it to be optimally processed, in terms of subsequent auto-correlation and integration in time and range. (The latency time of the buffer must therefore be long enough to allow multiple processing strategies to be applied before the low-level data are over-written.) The final optimally-derived data products, (which are typically at least an order of magnitude smaller,) are then transferred to the permanent data archive. At the same time, a second copy of the incoherent scatter data is separately passed through a default signal processing strategy in order to produce the quick-look data needed for control of experiments [6]. EISCAT-3D will provide visualisation means to present its data products and system status.

We consider a group of functions that support EISCAT-3D to collect raw data as a *data acquisition sub-system*; a group of functions that support of data storing and archiving as a *data curation sub-system;* a group of function that support data discovery and deliver to end users as a *data access sub-system*.

In the ***data acquisition sub-system***, EISCAT-3D collects 3 types of data [6]: 1) incoherent scatter data; 2) interferometric data, and 3) data from supporting instruments (which are not among EISCAT-3D main data products). We analyse the life-cycles of incoherent scatter data and the interferometric data as follows:

- Incoherent Scatter data are collected from the antenna elements, then beam-formed to generate the actual receiver beams. The data stream will be split and the two replicated streams of beam-formed data are handled in two different ways [7]. One of the streams will be directed to standard signal processing software or an analysis program to automatically correlate the data for some data analysis experiments. This will largely be an automated process using a default time integration strategy so that each experiment produces a continuous stream of auto-correlated data at a standard time resolution, with a standard gating strategy. The other stream of sample-level, beam-formed data will flow into a cyclic buffer where it can be stored for some finite time until it is over-written [7].
- Interferometric data are a set of incoherent scatter data with sub-beam width resolution. These data will be delivered to a separate interferometry buffer, where these data will be continuously tested against the coherency threshold criteria. If the threshold is not exceeded the data will be over-written; if the coherences on a given number of baselines begin to exceed the threshold, the content of the interferometry buffer will be continuously transferred to the central site ring buffer. Interferometry data will continue to be transferred to the ring buffer until the coherences have fallen persistently below threshold values for more than a user-specified period, providing ample time to examine the decay of the scatters. In this way, the raw data from the interferometry system can be stored in the central site ring buffer for long enough for it to be optimally processed into visibility patterns and brightness functions, which will then become part of the permanent archive [6]. Sufficient metadata will be generated to describe which data sets were used and how the interferometry calculation were carried out [8].

[7] provides full list of functions provided at this *sub-system*. The main request is a control software to provide the following computational functions:

- **Process Control**. E.g., a *Scheduler* is designed which is a control process that takes input from a *Time Allocation Committee* and any geophysical alert events and determines the current observing programme that should be executed. A *Control system* generates the low-level logic and drives the hardware and software components that do the actual work. The *Control system* is distributed over the various sites, whereas the *Scheduler* is located centrally;
- **Instrument Monitoring**, **Data Collection**, **(Parameter) Visualisation**, and **Instrument Calibration**. E.g., a monitoring system is responsible for collecting and collating all the auxiliary data that emanates from the EISCAT-3D radar system. This means all the engineering data that do not form part of the main data stream, but are essential for calibrating and interpreting that data. In addition, the data are presented to any *Operator* who is controlling the system;
- **Instrument Access**, and **(Real-time) (Data) Visualisation** E.g., an Operator is a combination of software and human interaction that drives the radar via the control system and makes tactical decisions regarding the operation of the radar experiment. The Operator can see incoming data via the Visualisation system in real-time and react appropriately to scientifically interesting events.
- **Temporary Data Storing**. E.g., the raw data is stored in the raw data buffer (store) for either reprocessing by a *Data selector* or for directly dumping it into a *Data archive*. A *Raw Data Accept* process accepts commands from the *Control system* and, when required, extracts data from the raw data buffer (store) and forwards it directly to the *Data archive*; [7]

- **Noise Reduction**. E.g., an *Integrator* time-integrates the signals and reduce the noise variance and the total data throughput of the system that reached the *Data archive* [7].

Here, the names of the functions are given as the abstractions of the requirements described in the EISCAT-3D design documents (or related information materials). The original requirements (from the EISCAT-3D design documents) are used as examples to illustrate the meanings of the functions. The same principle applies to the rest of the analysis. A more formal definitions of the functions will be provided in sub-section 3.7.

In the ***data curation sub-system***, EISCAT-3D will archive data delivered from the ring-buffer for the long term. The *Data Preservation and Distribution* component of the overall EISCAT-3D system is the one that essentially acts as an ingest facility, with provision for adaptive storage, data location management and the ability to reprocess data within certain time limitations. The data handled at this sub-system at typical time-associated data in file formats. The key functions include:

- **Fault Tolerant Data Buffering**. E.g., to allow the system to cope with anomalies in the regularity of the data flow. For instance, to allow a catch-up grace period for data which are temporarily disrupted. Such disruptions, in addition to the "standard operation" disruptions listed above, would also include more major incidents, such as rebooting a down-stream computer. The system would provide sufficient latency to recover, with time, the computer outage, as it would be possible to process the backlog of data faster than the incoming data rate. This would not apply in the case of the interferometric system;
- **Data Replication**. E.g., automated secure remote backup [8];
- **Data Preservation**. E.g., the data archive receives scientific data from the receiver signal processing streams  and auxiliary data from the monitoring system and serves it to users and the visualisation system.[7]

EISCAT-3D designs for the following functions for the ***data access sub-system***:

- **Data Discovery**. E.g., the data archive consists of *Store Query Process* that accepts commands from users and instructs the data stores to send data to a *Data Aggregator*. The *Data Aggregator* merges science and auxiliary data when required and converts the data into a format suitable for transmission to the user [7]
- **Data Access**. E.g., 'throughput intensive' mode of operation for access of data in file formats in high volumes is requested;
- **Data Visualisation**. E.g., EISCAT-3D will provide facilities for visualisation of metadata for event-search or maintenance purposes [8].
- **Access Control**. E.g., secure access for users [8];
- **Data Conversion**. E.g., for experimental purposes, the low-level format (raw bit streams) would need to be convert to more usable format, and transferred to conventional media. For example, automatic detection (and excision) of spurious signals, and the automatic recognition of scientifically interesting data may occur before post-integration [8].

EISCAT-3D will provide facilities for ***data processing***. The focus is on visualisation and the functions to be provided include [9]:

- **(Data) Visualisation** and **(Scientific) Visualisation**. Real-time visualisation of analysed data will be provided, e.g., with a figure of updating panels showing electron density, temperatures and ion velocity to those data for each beam. In addition, non-real-time (post-experiment) visualisation of the physical parameters of interest will also be provided, e.g.,

- by standard plots,
- using three-dimensional block to show to spatial variation (in the user selected cuts),
- using animations to show the temporal variation,
- allow the visualisation of 5 or higher dimensional data, e.g., using the 'cut up and stack' technique to reduce the dimensionality, that is take one or more independent coordinates as discrete; or volume rendering technique to display a 2D projection of a 3D discretely sampled data set.

To support its community, EISCAT-3D will provide the following functionality:

- **(Interactive) Visualisation**. E.g., EISCAT-3D wish to allow users to combine the information on several spectral features, e.g., by using colour coding, and to provide real-time visualisation facility to allow the users to link or plug in tailor-made data visualisation functions, and more importantly functions to signal for special observational conditions. [9]

EISCAT-3D introduces significant challenges in data handling that how to cope with large-scale data in real-time. The requirements are documented in detail and solutions are proposed. However, the system is not yet implemented, and there are many uncertainties in realisation. In the next, we will look at a more mature system, EURO-Argo.

## 3.2 Analysis of Euro-Argo

Argo is a global ocean observing system comprising of a large network of robotic floats distributed across the world's oceans and supporting infrastructure. It is a unique system to monitor heat and salt transport and storage,  ocean circulation and global overturning changes and to understand the ability of the ocean to absorb excess carbon dioxide from the atmosphere.

EURO-Argo[6] is the European contribution to Argo as an European Infrastructure. The objectives of the new Euro-Argo Research Infrastructure include 1) to provide, deploy and operate an array of around 800 floats contributing to the global array (a European contribution of ¼ of the global array); 2) to provide enhanced coverage in the European regional seas; and 3) to provide quality controlled data and access to the data sets and data products to the research (climate and oceanography) and operational oceanography (e.g. GMES Marine Core Service) communities. [10]

The robotic floats are operated as follows [11]: after being released, floats dive to a programmable depth (currently 1000 metres), drifting freely in currents. Every 10 days, a float dives to 2000 metres, then rises to the surface to send data by satellite link. More than 200 cycles can be performed during the float's 4 year lifespan. The data collected by Argo include heat, salt transport/storage, ocean circulation and global overturning changes in order to understand (amongst other things) the ocean's absorption of excess carbon dioxide.

The life-cycle of Argo data are as follows [12]:  the 11 national Data Assembly Centres (DACs) receives data from satellite operators, decode and perform quality control (according to a set of 19 real-time automatic tests). Erroneous data are flagged, corrected if possible and then passed to the 2 Global Data Assembly Centres (GDAC), and to the World Meteorological Office Global Telecommunication System (GTS). The 2 GDAC located at Coriolis (France) and USGODAE (USA) collect data from the 11 DACs and provide a unique access both in real time (within 24-48hrs after transmission) and delayed mode (6-12 months after transmission). Data available in NetCDF format in FTP and internet. The 2 GDACs synchronise every day. GDACs also deliver data to several Argo Regional Centres (ARCs), where the expertise on specific geographical ocean regions will provide comprehensive data sets (including non-Argo data). Data from GDACs will be long-term archived at data centre located in NODC (US).

The architecture of Euro-Argo is depicted in Figure 2.1 (D). We consider a group of functions that supports the 11 DACs to collect the raw data from the floats and standardise the collection process as a *data acquisition sub-system*; a group of functions that supports the 2 GDACs to check the data quality and to archive the data as a *data curation sub-system*; and a group of functions that support data distribution and access as a *data access sub-system*. EURO-Argo provides limited functions for *data processing* and *community support*. It links with external systems, such as MyOcean[7] (an ocean monitoring and forecasting system which provides products and services for all marine applications) and the SeaDataNet[8] (a Pan-European Infrastructure for Ocean & Marine Data Management), to provide such functionalities.

In the ***data acquisition sub-system***, EURO-Argo includes the following functions [11]:

- **Data Collection**. E.g., to receive data from satellite;
- **Data Conversion**. E.g., to convert the standard exchange formats;
- **Noise Reduction**. E.g., to apply standardised real-time quality control;
- **Data Transmission**. E.g., delivering data through FTP to the GTS and GDACs within 24hrs of surfacing and to principle investigators (PIs) on a more relaxed schedule;
- **Process Control**. E.g., to allow the coordination of Argo data handling for the floats under their control.

In the ***data curation sub-system***, the 2 GDACs keep the master copies of the Argo global dataset (metadata, profiles, trajectories and technical information). The data are long-term archived at the NODC centre in US. The key functions provided include:

- **Replica Synchronisation**. E.g., storages of the 2 GDACs centres are synchronised everyday [11];
- **Data Preservation**. E.g., NODC will long-term archive all Argo data [11];
- **Data Quality Verification**. E.g., the Argo Regional Data Centres receive data from the GDACs and look at data from ocean basins to verify the consistency of float data and generate products. Feedback will send to PIs [11];
- **Data Product Generation**. E.g., comprehensive data sets (including non-Argo data) are provided by the Argo Regional Data Centres [11].

In the ***data access sub-system***, EURO-Argo provides the following functions:

- **Data Publication**. E.g., GDACs distributes data to users, the Argo Information Centre, the NODC archive centre, and the RDACs via FTP and the internet. [11]

[7] MyOcean: http://www.myocean.eu.org/

[8] SeaDataNet: http://www.seadatanet.org/

The EURO-Argo system is relatively mature and capable of supporting the whole life-cycle of Argo data from acquisition to preservation. However, only necessary and basic operations are provided. Many processes have not yet been automated and comprehensive functionalities, in particular for data access and data processing, have not yet been considered. Compared to EURO-Argo, ICOS offers more implementation experiences in such functions.

## 3.3 Analysis of ICOS

ICOS[1], the Integrated Carbon Observing System, is a world-class research infrastructure to quantify and understand greenhouse gas fluxes. The objectives of ICOS community are to monitor greenhouse gases (GHG) over the long term through atmospheric, ecosystem and ocean networks.

As shown in Figure 2.1 (A), the ICOS distributed infrastructure consists of the following elements [13]:

- **Headquarters**, which coordinates the research infrastructure at the European level. This mainly involves human activities;
- **The ICOS network** of atmospheric, ecosystem and ocean observation sites, which include 50 atmospheric, 50 ecosystem, and 10 ocean stations;
- **A Central Analytical Laboratory** for calibration, and air samples analyses including radiocarbon for the entire network;
- **An Atmospheric Thematic Centre** (ATC) which is responsible for the coordination of atmospheric measurements, instrument development/servicing, and online data processing;
- **An Ecosystem Thematic Centre** (ETC) which is responsible for total ecosystem flux measurements and component fluxes and carbon pools, including data processing and instrument development; and
- **An Ocean Thematic Centre** (OTC) which is responsible for coordinating continuous marine observations, initial data processing from marine network.

We consider a group of computational functions that facilitates the collections of the observations of greenhouse gases (GHG) from the hundred plus stations of ICOS atmospheric, ecosystem and ocean networks as a ***data acquisition sub-system***. The design [13] describes the requirements for a control software which offers the following functions:

- **Instrument Access**. E.g., to provide the access and control of the stations locally and remotely [13];
- **(Parameter) Visualisation**. E.g., to display the parameters and measured variables in real-time [13];
- **Instrument Configuration**. E.g., to allow the configuration of the stations [13];
- **Instrument Configuration Logging** and **Instrument Monitoring**. E.g., to log the state of the station configuration and possible problems (warnings and alarms), and to implement the security routines [13];
- **Process Control**. E.g., to include a central sequencer for all stations instruments; and to allow performing action on all controllable components of the stations, to permit the execution of measurement sequences, and to allow the handling of the data flow and transfer [13];
- **Data Collection** and **Instrument Calibration**. E.g., to automate the operation such as, measurement, and calibration [13];

- **Message Handling**. E.g., to allow the handling of the intra- and extra- station communication [13];.
- **Instrument Integration**. E.g., to permit the addition of further instruments to the station in the future [13];
- **Data Transmission** and **Data Transmission Monitoring**. E.g., to allow transferring data from all ICOS stations in real-time. A protocol will be set up to check for the correct transfer of the expected daily data files. Automatic emails will be sent to the stations principle Investigators (PIs), in case of problem in the data-transmission/gathering process [13];
- **Noise Reduction**. E.g., quality check will be applied based on the information given by each instrument (e.g. temperature, flow rates), and on a statistical algorithm to identify suspicious signals.

We consider a group functions that support the ICOS three thematic centres, ATC, ETC and OTC, to receive observations from stations, check data quality and to archive data as the ***data curation sub-system***. ICOS designed or implemented the following functions in this sub-system:

- **Data Quality Verification**. E.g., interactive tools are provided to the station PIs in order to flag the data which correspond to malfunctions of the instruments or to local contaminations [13];
- **Data Preservation**. E.g., the raw data received from stations will be archived. Metadata about the sites, the variables and the instruments is provided when archiving data. The data archiving will be dynamic, complete and robust over time (20 years) to allow for an automatic reprocessing of the whole dataset for instance when primary scale changes will occur (every 3 years on average) [13];
- **Data Product Generation**. E.g., based on Level 0 (raw data), ATC will produce Level 1 data which are expressed in geophysical units and usable by modellers to calculate fluxes, and Level 2 data which are in time series [13];
- **Data Versioning**, which tracks changes in data and preserves different versions of data products. E.g.,level 1, 2 data products will be organised into versions, resulting from the regular re-processing of level 0 data (raw data) [13];
- **Data Identification**. E.g., the implementation of a dataset ID system for ICOS data, such as a digital object identifier (DOI) is important for tracking and referencing data resources (e.g. http://datacite.org/). Attaching a DOI to a data set will be achieved by “freezing” the database at regular intervals (at least annually) and feeding data streams to the ICSU World Data System [13].

A group of functions which supports the publication and access of ICOS data products is considered as a ***data access sub-system***. ICOS designs for a Carbon Portal to distribute its data products. For example, the ATC has implemented the following functions in its web portal[9]:

- **(Data) Visualisation**. E.g., graphic products are provided including the time series data plots of the measurements and daily automated diagnostic plots for station and instruments monitoring. A simple PHP browser with thumbnail capacity is available [14];
- **Data Publication**. E.g., the ATC provides on-request pipes, and hourly carbon dioxide or methane measurements are made available by SFTP server, or HTTPs server [14].

The Carbon Portal will also support discovery and integration of the data from its 3 thematic centres, and the following functionalities are planned [13]:

[9] ICOS web portal: https://icos-atc-demo.lsce.ipsl.fr

- **(Data) Annotation** and **Data Publication**. E.g., 1) interface for metadata description and data release publications, and 2) web-service for external data access;
- **Data Discovery**. E.g., the Carbon Portal will generate and provide effective tools to discover, find, extract and collocate observations according to user needs;
- **Semantic Harmonisation** and **Data Conversion**. E.g., the harmonisation of data and metadata standards, together with graphical formats and links to new products, will be coordinated between the TCs, the Carbon Portal and the station principal investigators. The Carbon Portal will offer different options to meet user needs with online/offline automatic conversion tools.

We consider a group of computational functions that supports analysing and mining of ICOS data as a ***data processing sub-system***. ICOS will provide the following functions through its Carbon Portal [13]:

- **Data Mining** and **Data Extraction**. E.g., ICOS Carbon Portal will provide interfaces for data mining and data extraction;
- **Scientific Modelling**. E.g., ICOS will provide Level-3 data. The generation of Level-3 data is a research process, and it is desirable in this process that several models are applied to the same Level-1 or Level-2 data, in order to obtain ensemble estimates of Level-3 data using models prescribed with different parameter values and/or of different structure. This is similar to IPCC's combination of results from several climate models forced by the same scenario, hence giving a range of uncertainty on a model's projections.
- **Data Assimilation**. E.g., ICOS will provide access to quality-controlled long-term observational data and data products for data assimilation and modelling.

ICOS designs the following tools/functions to be provided by the Carbon Portal to support the ICOS user community. This group of functions can be considered as a ***community support sub-system***.

- **Data Citation Tracking**. E.g., Processed and quality controlled dataset offered via the Carbon Portal will be frozen on a yearly or semi-annual basis and be published in specialized geosciences data journals. Citation and referencing of such papers will offer a simple bibliometric means of tracing and measuring the data usage through its referencing in the scientific literature. The Carbon Portal will keep track of published scientific papers using ICOS data and provides links to them. Any other outcome of the use of the ICOS data will also be documented on the Carbon Portal. Relevant information on data usage (and ICOS visibility) will be collected [13];
- **Authentication**. E.g., the Carbon Portal will develop an overarching registration system for all ICOS data streams and implement for itself and the thematic centres single-ID concepts (e.g. OpenID or GEO-ID) that allow logging in with one ID at multiple sites. The log-in procedure should be minimally intrusive and require only the acceptance of the data use policy (license), indication of purpose and affiliation [13];
- **Data Description Publication**. E.g., the Carbon Portal will coordinate with the thematic centres on the peer-reviewed publication of descriptions of the ensemble of the databases. This publication ensures bibliometric recognition of principal investigators' and thematic centres' work [13].

Besides the atmospheric thematic centre, ICOS is constructing the similar information systems for the other two thematic centres. However, in a long-term, ICOS encounters the challenges of aggregating and integrating data across the 3 thematic centres and to conduct scientific analysis and experiments upon the integrated data. In such areas, EMSO is a step ahead.

## 3.4 Analysis of EMSO

EMSO[3], European Multidisciplinary Seafloor Observatory, is a European network of sea floor observatories for the long-term monitoring of environmental processes related to ecosystems, climate change and geo-hazards. The objectives of EMSO community are to ensure the technological and scientific framework for the investigation of the environmental processes related to the interaction between the geosphere, biosphere, and hydrosphere and for a sustainable management by long-term monitoring also with real-time data transmission.

EMSO observatories will include a common set of sensors for basic measurements and further sensors for specific purposes defined by users. The common set of instruments comprises seismometers, hydrophones for geophysics, magnetometers, gravity meters, CTD (Conductivity, Temperature, and Depth), current meters, chemical sensors, pressure sensors, and hydrophones for bio-acoustic monitoring. Additionally, laboratory studies are performed on material collected at these sites by sampling devices (e.g., water samplers, sediment cores, traps etc.). The following activities are carried out at EMSO individual observatories [15]. They are likely supported by computational facilities in order to:

- Design measurements and monitoring models based on geographical location, scientific requirements, operational requirements, and available resources;
- Develop and test the sensor detectors;
- Deploy the instruments into selected locations of the deep-sea;
- Adapt the infrastructure to new deployed instruments. EMSO-ERIC Scientific and Technological Advisory Board and the Executive Board will establish general and detailed requirements and standard, in order to fulfil both cabled node and stand-alone node integration into a unique research infrastructure;
- Recover and reset the sensors;
- Update with new technology e.g., using cabled observatory;
- Send data from sensors to surface buoys/boats, to satellites then forward to shore stations.

EMSO data collected in experiments at 11 regional sites are locally stored and organized in catalogues or relational database and run by the institutions involved. Some of EMSO observatories' data from some distributed sites are harvested and long term archived at 3 data archives, Ifremer(EUROSITES[10]), UniHB(PANGAEA[11]) and INGV(MOIST[12]). A central archive hosting a web-service access to all the databases is planned for the near future. We consider a group of functions provided by MOIST, PANGEA and EUROSITES that support data quality control and preservation as a ***data curation sub-system***. Key functions include but are not limited to:

- **Data Identification**. E.g., assigning persistent identifiers. EMSO partially already assigns DOIs to some of its data sets, e.g., from the HAUSGARTEN site, and will use DOIs for further data products;
- **Data Cataloguing**. E.g., EMSO collects metadata on both the physical sensors and observatories as well as on the data. Observatories are intended to be described by SensorML;
    - Metadata is provided by the regional nodes of EMSO: MOIST[12] is a data management system for multi-parametric observatories, aiming at hosting multidisciplinary data and

[10] EUROSITES: http://www.eurosites.info/about.php

[11] PANGAEA: http://www.pangaea.de/

[12] MOIST: http://moist.rm.ingv.it/

metadata. The core part is the database that indexes data and keep track of the data source. MOIST supports 11 EMSO sites in organising, indexing and transforming data into a compatible data scheme. MOIST is developed to adopt the most common standards (e.g., OGC, NASA, INSPIRE) for organising its information system. It offers access to data in e.g. NASA DIF and Dublin Core via a OAI-PMH interface as well as an OpenSearch interface. The information system PANGAEA offers a variety of services for scientific project data management, long-term data archiving and data publication. PANGAEA provides access to metadata in several formats such as DC, NASA DIF, ISO19139 or DarwinCore. Access to metadata is provided via a OAI-PMH interface, OpenSearch as well as DiGIR. Ifremer offers access to several EMSO sites via their EUROSITES data management system, which offers access to data in NetCDF format. Via FTP.

  - EMSO has implemented a prototype common data catalogue[13] which uses the OpenSource panFMP[14] software to harvest and index the metadata records. PANGAEA and MOIST metadata are harvested via their OAI-PMH interfaces while for Ifremer an additional service has been implemented which extracts metadata from the NetCDF records which then are harvested via http. The EMSO common data catalogue offers a common OpenSearch interface as well as a metadata transformation service which offers metadata in dclite4g format compliant with the GENESI requirements used for the ENVRI OpenSearch client;

- **Data Preservation**. E.g., in PANGAEA,  a curator is responsible for the data archiving and publication[15]. MOIST will adopt a reference model developed inside SCIDIP-ES EC project.

The PANGAEA data library and publisher retrieves data from EMSO resources and make them publically accessible. We consider a group of functions that facilitates the publication and access of EMSO data as a ***data access sub-system***. This sub-system includes the following functions:

- **Data Conversion**. E.g., PANGAEA provides the following tools/software:
  - Pan2Applic[16], which converts files or folders of files (ascii/tab-separated data files with or without metaheader), downloaded from PANGAEA via the search engine or the data warehouse to formats as used by applications, e.g. for visualization or further processing;
  - PanTool[17] which is used for data conversion and recalculation, written to harmonize individual data collections to a standard import format used by PANGAEA;
  - Split2Events[18], which splits one file with data from several events into several files, one for each event;
  - PANGAEA as well as MOIST are planning to provide NetCDF transformation services.
- **Data Visualisation**. E.g., PANGAEA provides the following tools/software:
  - PanPlot[19], which allows the visualisation of data versus time or space in standard x-y-plots or ternary diagrams;
  - PanMap[20], which is for the geographical presentation of data in maps;

[13] EMSO common data catalogue: http://dataportals.pangaea.de/emso

[14] panFMP: www.panfmp.org

[15] PANGAEA data curation and management: http://wiki.pangaea.de/wiki/Project_data_management

[16] PANGAEA Pan2Applic: http://wiki.pangaea.de/wiki/Pan2Applic

[17] PANGAEA PanTool: http://wiki.pangaea.de/wiki/PanTool

[18] PANGAEA Split2Events: http://wiki.pangaea.de/wiki/Split2Events

[19] PANGAEA PanPlot: http://wiki.pangaea.de/wiki/PanPlot

  - ○ GIS[21]. Visualisation of geo-referenced data in PANGAEA through GIS (Geographical Information System) functionality is enabled by using Google Earth;

  MOIST provides the following tools/software:

  - ○ MOIST Plot, a multi-parametric plot in an interactive area of an online web page, for a first immediate data analysis by the user, before download selected data.

- **Data Publication**. E.g., PANGAEA offers a DOI resolution service and makes internally use of a DOI registration service to register the DOI metadata records at DataCite. In cooperation with publishers such as Elsevier PANGAEA provides a cross linking service which allows 'reciprocal linking' – automatically linking research data sets deposited at PANGAEA to corresponding articles in Elsevier journals on its electronic platform ScienceDirect and vice versa;
- **Data Import**. E.g., Data import tools[22] are provided by PANGAEA;
- **Data Discovery**. E.g., a Google-like advanced metadata discovery is provided for public access[23]; also the Common EMSO data catalogue and data portal;
- **Data Citation**. E.g., PANGAEA supports for data citation[24]. Each data point is fully citable with a DOI, and can be cross-reference with journal articles. It also supports of pre-publication, peer-review process and support of data citation;
- **Quality Verification**. E.g., PANGAEA's data policy[25] describes the quality assurance for data submission. QC procedures are maintained within the PANGAEA data curatorial process[26] during which quality flags can be assigned to indicate the quality of each measurement;
- **Metadata Harvesting**. E.g.,
  - ○ PANGAEA OAI-PMH for ESONET data in EMSO sites: harvesting test, integration into ENVRI metadata catalogue etc.;
  - ○ PANGAEA GeoRSS : Embedding GeoRSS feed;
  - ○ Ifremer SOS (Sensor Observation Service) for EUROSITES oceanographic data in EMSO sites: getCapabilities, getObservation, check O&M format;
  - ○ PANGAEA SOS for INGV data in EMSO sites (via MOIST: moist.rm.ingv.it): getCapabilities, getObservation, check O&M format;
  - ○ MOIST OpenSearch for INGV data and metadata in EMSO sites: Data and metadata search according to time or space or parameter;
  - ○ Common NetCDF metadata extraction and transformation service;
  - ○ MOIST OAI-PMH for harvesting INGV data and metadata in EMSO sites.
- **(Identifier) Registration**. E.g., a catalogue of EMSO DOIs is being planned in order to register EMSO DOIs;
- **(Metadata) Registration**. E.g., a centralised Metadata Catalogue to store metadata harvested from distributed sites is implemented within the prototype EMSO data catalogue and data portal (see above);

[20] PANGAEA PanMap: http://wiki.pangaea.de/wiki/PanMap

[21] PANGAEA GIS: http://wiki.pangaea.de/wiki/GIS

[22] PANGAEA data import tool: http://wiki.pangaea.de/wiki/Import

[23] PANGAEA advanced metadata discovery: http://www.pangaea.de/

[24] PANGAEA data citation: http://wiki.pangaea.de/wiki/Citation

[25] PANGAEA data policy: http://wiki.pangaea.de/wiki/Data_policy

[26] PANGAEA data quality control: http://wiki.pangaea.de/wiki/Data_policy#Quality_assurance

- **(Sensor) Registration**. E.g., EMSO aims to implement core standards of the Open Geospatial Consortium (OGC) Sensor Web Enablement (SWE) suite of standards, namely the OGC standards SensorML, Sensor Registry, Catalogue Service for Web (CS-W), Sensor Observation Service (SOS) and Observations and Measurements (O&M). A sensor registry is available at ESONET[27].

We consider a group of functions that support EMSO users to conducts various tasks as a ***community support sub-system***. We identified the following functions:

- **Accounting**. E.g., the statistics of the portal accesses is planned to replace the user registration, which can track resource consumption by users for the purpose of capacity and trend analysis;
- **Metadata Submission**. E.g., PANGAEA provides online metadata & data submission service[28], which supports metadata standards such as ISO19115, Dublin Core, DIF, DarwinCore, and DataCite metadata;
- **Curation Editor**. E.g., PANGAEA also provides online curation editor[29] to be used by curators for the administration of metadata and the import of data;
- **Event Notification**. E.g., real-time access of the sensor data to identify new phenomena, and events occurring to provide geo-hazard warning [15].

EMSO provides advanced technology in data publication and citation through the PANGAEA system. EMSO also offers capabilities for data access, standardisation/harmonisation and visualisation via MOIST data infrastructure. Presently (in Dec. 2012), 3 regional sites data are integrated in MOIST, and one regional site is integrated in PANGAEA which additionally offers data from several related or preparatory studies for other EMSO sites. In addition, Ifremer offers access to data from all EUROSITES sites which are shared with EMSO. EMSO has integrated all its operational sites within a common data portal. In the next step, EMSO plans to continue to harmonize its vocabularies and terminologies according to SEADATANET standards and aims to offer access to data via a common NetCDF format which is compliant with SEADATANET. Further EMSO plans to improve standardised access to real time data via SOS.

In the next, we look at EPOS, which has special emphasis on the integration and interoperability problem and tackles the problem by a new infrastructure design.

## 3.5 Analysis of EPOS

EPOS[2], the European Plate Observing System, is a research infrastructure and e-Science for data and observatories on earthquakes, volcanoes, surface dynamics and tectonics. The objectives of EPOS community are to integrate the existing research infrastructures (RIs) in solid Earth science in order to increase the accessibility and usability of multidisciplinary data from seismic and geodetic monitoring networks, volcano observatories, laboratory experiments and computational simulations. EPOS aims to enhance worldwide interoperability in Earth Science by establishing a leading integrated European infrastructure and services.

Since only the seismic network of EPOS is relatively mature when writing this report in Dec. 2012, the following analysis is limited to the requirements of this discipline.

[27] ESONET: http://vps.dbscale.com:8080/esonet/

[28] PANGAEA online metadata & data submission service: http://www.pangaea.de/submit/

[29] PANGAEA online curation editor: http://wiki.pangaea.de/wiki/4D

EPOS focuses on integration and interoperability of existing earth science systems. It does not actively design or implement functionalities for ***data acquisition*** and ***curation***. Such functionalities are already available in the existing systems. For example, the real-time seismic waveform data from more than 500 broadband stations in Europe are collected by the Virtual European Broadband Seismograph Network (VEBSN), using seismic data acquisition systems such as, Antelope, SeiscomP/SeedLink, and SCREAM [16]. A number of data centres, such as ORFEUS and EMSC, respond to data quality control and archiving. Data are archived using archive protocols (e.g., ArcLink and mseed2dmc). All data is openly available to the research community through a variety of means, such as web services, direct access and interactive tools. In the long term, the data will be preserved via EUDAT nodes using grid data technology such as iRODS, which store and replicate the data, providing also unique and persistent ID (PID) to data granules through a federated handle systems [17].

We consider a group of computational functions provided by VEBSN to support data collection as a ***data acquisition sub-system***. The key functions include:

- **(Real-time) Data Collection**. E.g., the VEBSN facilitates real-time collection, locations and quantifications of important seismic events waveform data [16];
- **(Real-time) Data Transmission**. E.g., VEBSN data are exported in real-time by Seedlink protocol. VEBSN facilitates rapid centralized data exchange between European observatories and the IRIS-DMC data centre [16];
- **(Real-time) Data Extraction**. E.g., the VEBSN also provides rapid accurate locations of large to medium sized earthquakes, complete with automatic picks and magnitude estimates, to the EMSC and observatories in the region. EMSC combines the VEBSN picks with additional data picks to produce an improved location and magnitude estimate with a certain delay. The event waveform data are also used for routine rapid moment tensor determinations [16].

We consider a group of computational functions provided by ORFEUS to support data quality control and archiving as a ***data curation sub-system***. The key functions include:

- **Quality Checking**. E.g., ORFEUS processes raw data according to various quality parameters. The continuous seismic waveform data from the Virtual European Broadband Seismograph Network (VEBSN) are monitored at ORFEUS Data Center (ODC) in various ways to ensure a high quality of waveform data and metadata. The following monitors are in place [18]:
  - Near real-time monitor of the Power Spectral Density (PSD) versus time for selected frequencies;
  - Network Latency Monitor--monitors latency per network and 3 days history;
  - Near real-time monitor of the PSD through a Probability Density Function (PDF);
  - Histograms for magnitude;
  - Time residuals; and
  - qc-plots.
- **Data Identification**. E.g., assignment of persistent identifiers (PIDs) to data collections [19];
- **Data Cataloguing**. E.g., to associate metadata to data collection, and update metadata when associated data collections change [20];
- **Data Preservation**. E.g., repositories for observational and experimental raw data, pre-processed data products and modelling/simulation data including the respective metadata [16];
- **Data Replication**. E.g., EPOS distinguishes safe replication which supports bit-stream preservation, optimal data curation and accessibility [19];

- **Data Staging**. E.g., dynamic replication which is the replication of data between storage resources and HPC staging areas [19];
- **Workflow Enactment**. E.g., to automate a set of operations [19];
- **Provenance Tracking**. E.g., to track data sources and processes [19].

Non-functional requirements emphasise on performance aspects including, security, consistency, productivity, responsibility, reliability, accessibility, availability, scalability, and load-balance.

EIDA [21] serves in the EPOS data infrastructure as a consortium of waveform data centres that share a common agreement on issues related to data formats, metadata, transfer protocols and interfaces within the consortium. We consider a group of functions provided by the EIDA data centre which supports of data exchange and discovery as a ***data access sub-system***. The technical architecture consists of the ArcLink middleware, which is installed at each node of the consortium. Each node synchronises its network, station, location, channel metadata everyday [21]. On top of ArcLink each node has built its own infrastructure to exchange the waveform data within the consortium through the TCP/IP protocol [21]. This presents peer-to-peer communication. Federated security is being planned for each individual institution within EPOS, so that each institute can maintain its own security infrastructure, but a single sign on process is desired, probably making use of some combination of X509 certificates, Shibboleth and LDAP in order to make an apparently seamless AAI (Authentication and Authorisation Infrastructure). [17]

To summarise, the functions and embedded computations provided by the ***data access sub-system*** include:

- **Data Publication**. E.g., data and metadata associated are made publically accessible by data centres [22];
  - Web-services which provide access by stand-alone clients to download bulk-data using command line/batch systems;
  - E-mail based data request services, such as NetDC, BrequFast and AutoDRM.
- **Data Transmission**. E.g., network protocols for moving large or small amounts of data, and for moving real-time or non-real-time data;
- **Data Discovery**. E.g. ArcLink, a technology that is used to query a server for seismological data in a certain time window and region, which can handle requests of metadata, waveform, quality control and routing [21];
- **Access Control**. E.g., externally, expert users will likely be permitted to interact with resources via the command line using standard Grid credentials [17].

For ***data processing***, EPOS data centres, such as ORFEUS, have established long lasting tradition for data analysis and mining. ORFEUS maintains a repository [23] of software/tools for specific interest to the seismological community with emphasis on free software. The required functions for data processing mainly include the following areas:

- **Data Conversion**. E.g., to convert waveform formats [23];
- **Data Analysis**. E.g. for routine analysis of signals or hazard analysis [23];
- **(Data) Visualisation**. E.g., statistical plots [23];
- **Scientific Simulation** and **Scientific Modelling**. E.g., provide simulation and modelling of solid stress and strain due to elastic static response to an earthquake [23]; geomechanic modelling, and wave propagation modelling [23];

- **(Scientific) Visualisation**. E.g., visualisation and analysis of seismograms [23].

The functionality provided by EPOS to support its community, which can be considered as a ***community support sub-system***. Such functionality includes but is not limited to:

- **Authentication**. E.g., Federated security mechanism, by using Shibboleth-based AAI or OpenID for web-based access [19];
- **(Interactive) Visualisation**. E.g., ORFEUS provides the following stand-alone services via web clients [24]:
  - Event selection from catalogues based on user defined regions and/or magnitude thresholds, such as Wilber II;
  - Stream selection on network, station and channel level, based on geographical region and epicentral distance, such as OWI;
  - Direct and automatic waveform harvesting (SEED) from EIDA;
  - Time window adjustment using configurable phase arrival times.
- **Event Notification**. E.g., the EMSC operates an Earthquake Notification Service[30] which implements a software named QWIDS (Quake Watch Information Distribution System) to provide a quick and robust data exchange system through permanent TCP connections. QWIDS disseminates email/SMS/fax to the registered end-users within 20-30 minutes on average after the earthquake occurrence.

EPOS designs for a new-generation earth science system by applying the most advanced e-Science technologies on existing well-developed (seismic and other earth science) systems. However, the project is still in its early stages and design work is not yet completed. In the next section, we will look at LifeWatch, which addresses similar challenges to EPOS in many areas, and has provided some solutions through its own design study.

## 3.6 Analysis of LifeWatch

LifeWatch[5] is an e-science and technology Infrastructure for biodiversity and ecosystem research to support the scientific community and other users in the public, commercial, and policy sectors. The main objective of LifeWatch is to put in place the essential infrastructure and information systems necessary to provide an analytical platform for the use of both existing and new data on biodiversity. Different from an observatory system, such as EISCAT-3D or EURO-Argo, LifeWatch is an comprehensive integration infrastructure for domain-specific scientific data and computation. The emphasis is on a network of services providing secure access across multiple organisations to biodiversity and related data and to relevant analytical and modelling tools to collaborative groups of researchers [25].

The guidelines for the specification and implementation of the LifeWatch ICT infrastructure is given by the LifeWatch Reference Model [25], which is built on the ORCHESTRA Reference Model, an architectural framework for distributed processing and geospatial computing, which itself is based on ODP.

The Lifewathc Reference Model describes the LifeWatch architecture which consists of 4 function domains. As shown in Figure 2.1 (B), they the Resource Layer, the Infrastructure Layer, the

[30] EMSC Earthquake Notification Service: http://www.emsc-csem.org/Earthquake/seismicity/real_time.php

Composition Layer, and the User Layer [25]. The Resource Layer contains the data from sites and collections, but also contains catalogue services, analysis tools and processing resources that already exist at external networks; the Infrastructure Layer provides mechanisms for uniform access and integration of heterogeneous resources in the Resource Layer. Functional components in the LifeWatch Infrastructure Layer are implemented as services; the Composition Layer provides the tools for intelligent selection and orchestration of services, including workflows, semantic metadata for the discovery of components and the storage of additional attributes such as provenance and version information; and the User layer provides domain-specific presentation environments and tools for community collaborations, which is a generic portal with extended domain- and application- specific portlets.

With the ***data acquisition sub-system*** being absent, the 4 functional domains defined in LifeWatch are approximate to the 4 common sub-systems identified in ENVRI. The mapping is provided in Table 3.1.

Table 3.1: LifeWatch Functional Domains via ENVRI Common Sub-Systems

| **LifeWatch Functional Domains** | **ENVRI Common Sub-systems** |
|---|---|
| Resource | Data Curation |
| Infrastructure (Data Access & Discovery & Semantic Mediation) | Data Access |
| Infrastructure (Data Process & analysis & Modelling) + Composition | Data Processing |
| User | Community Support |

[25] provides a list of services to be provided by LifeWatch. Unfortunately, the mapping between these services to the 4 LifeWatch architectural layers are missing from the document. For the purpose of analysis, we examine the specification of each service and distribute them into the appropriate sub-systems.

In the ***data curation sub-system***, the LifeWatch consider those data, processing tools and instruments managed by multiple organisations, and in general, LifeWatch cannot dictate their location or configuration. The main function provided from within LifeWatch is a group of source integration services (Service Delegation) which provide an encapsulation of external resources to be used by the infrastructure. [25]

In the ***data access sub-system***, LifeWatch provides the following functions for data discovery and access in particular upon the heterogeneous resources [25]:

- **(Resources) Annotation**. E.g., an *Annotation Service*, which automatically generation of specific meta-information from various sources and relation with semantic descriptions;
- **Data Publication**. E.g., a *Catalogue Service*, which supports the ability to publish, query, and retrieve descriptive information for resources data, independent of a specific meta-information standard;
- **Data Discovery** and **Data Retrieval**. E.g., a *Feature Access Service*, which allows interoperable read and write access on feature instances available in an Service Network; a *Document Access Service*, which is a specialization of the *Feature Access Service* supporting access without manipulation to documents of any type; a *Taxonomy Access Service*, which is a specialization of *Feature Access Service* that allows to read (and write) taxonomic information; a *Provenance*

*Service*, which is a specialization of the *Feature Access Service* for accessing provenance data generated through annotation services, among others; and a *Data Discovery Sensor Access Service*, which is a specialization of the *Feature Access Service* for accessing sensor data, configuring a sensor and publishing sensor data;

- **Semantic Harmonisation**. E.g., a *Thesaurus Access Service*, which provides read and write access to a (multi-lingual) thesaurus for the vocabulary used on a service network; a *Ontology Access Service*, which provides read access to the specification of a logical ontology, and export or import of complete specifications into an ontology store; and a *Schema Mapping Service*, which is the mapping of features into a target schema through the transformation of each data instance from one data structure into another one preserving the original meaning;
- **Data Conversion**. E.g., a *Format Conversion Service*, which allows the conversion of data given in one format to the corresponding data given in another format; a *Coordinate Operation Service*, which changes coordinates from feature locations from one coordinate reference system into another; and a *Calendar Service*, which provides the transformation, comparison and arithmetical operations on data/time functions;
- **Data Compression**. E.g., a *Compression Service*, which performs data compression.

The services provided at the ***data processing sub-system*** include [25]:

- **Service Naming**. E.g., a *Name Service*, which encapsulates the implemented name policy for service instances in a service network;
- **Scientific Workflow Enactment**. E.g., a *Process Service Chain Access Service*, which supports the creation of an executable service instance based on an explicit description of a service chain; and a *Workflow Enactment Service*, which is a specialization of processing service, that allows the execution and monitoring of a workflow or a service chain;
- **Data Processing**. E.g., a *Processing Service*, which is a common interface for services offering processing operations by initiating the calculation and managing the outputs to be returned to the client;
- **Process Monitoring**. E.g., a *Service Monitoring Service*, which provides an overview about *Service Instances* currently running within a *Service Network*;
- **Scientific Modelling**. E.g., a *Modelling Services*, which is specialization of processing services, allows the user to discover, specify input for, and control execution of a variety of models (e.g. for simulation);
- **(Scientific) Visualisation**. E.g., a *Portrayal Service* (Map and Diagram Service), which visualizes, symbolizes, and enables geographic clients to interactively visualise geographic and statistic data by providing a graphical representation of the data;
- **Data Association**. E.g., a *Geolinking Service*, which allows establishing a virtual join between data having a spatial location and data without spatial location but referring to the same feature through common properties; a *Geocoder Service*, which allows adding geographic information to address data; and a *Gazetteer Service*, which allows to relate a geographic location instance identified by geographic names with an instance identified by coordinates;
- **Data Extraction**. E.g., a *Generalisation Service*, which allows to create spatial, temporal and other generalisation of features according to a given hierarchy; and an *Occurrence Distribution Service*, which allows creating distribution maps from a particular specie or specie group.

Finally, the ***community support sub-system*** provides the following functionalities [25]:

- **Authentication**. E.g., an *Authentication Service*, which verifies genuineness of principals using a set of given credentials;
- **Authorisation**. E.g., an *Authorisation Service*, which gives a compliance value as response to a given authorisation context;
- **User Registration**. E.g., a *User Management Service*, which creates and maintain subjects including groups of principals as entities that need authentication;
- **Instant messaging**. E.g., a *Communication Service*, which provides the harmonized access to direct user-to-user communication means based on multi-media technology and data-exchange between users (like chat, teleconference, SMS);
- **Template Generation**. E.g., a *Reporting Service*, which creates reports using actual information from other services according to a template (wrapper interface for existing products);
- **Data Editor**. E.g., an *Interpolation Service*, which allows interpolation of spatial locations;
- **Event Notification**. E.g., a *Notification Service*, which allows the sending of messages to a client, which previously has been registered to listen for certain events.

LifeWatch investigated the possibility of integrating various state-of-the-art standardised technologies to provide generic services and operations to support biodiversity research. This, on the other hand, results in high-level of abstractions of the design, which is likely to introduce difficulties in interpretation and realisation.

## 3.7 The Common Functions and Embedded Computations

To summarise the above observations, Table 3.2 lists functions and embedded computations provided by the existing research infrastructures. Each function is defined as an **interface** which encapsulates a set of required operations or services that act upon an *object*. Recall the definition of an **object** in ODP, which is a model of a real-world entity, characterised by its behaviour and its state. The interactions that occur between those objects at their *interfaces*.

The value domain used in the table is defined as follows:

$V$ ={Yes, No, Unknown, Not Applicable, In consideration}, where

- Yes: Evidences have been found that the research infrastructure provides the specified function.
- No: Evidences have been found that the research infrastructure does not provide the specified function.
- Unknown: Evidences haven't been found whether the research infrastructure provides the specified function.
- Not Applicable: The specified function is out of scope of the research infrastructure planned work.
- In consideration: Evidences have been found that the research infrastructure is in consideration of providing the specified function.

Table 3.2: The Common Functions and Embedded Computations

| A | Data Acquisition Subsystem | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **No** | **Functions** | **Definitions** | **ICOS** | **EPOS** | **EMSO** | **EISCAT-3D** | **LifeWatch** | **EURO-Argo** |
| A.1 | Instrument Integration | An interface that provides operations to create, edit and delete a sensor. | Yes | Unknown | Yes | No | Not Applicable | Yes |
| A.2 | Instrument Configuration | An interface that provides operations to set-up a sensor or a sensor network. | Yes | Unknown | Yes | Yes | In Consideration | Yes |
| A.3 | Instrument Calibration | An interface that provides operations to control and record the process of aligning or testing a sensor against dependable standards or specified verification processes. | Yes | Unknown | Yes | Yes | In Consideration | Yes |
| A.4 | Instrument Access | An interface that provides operations to read and/or update the state of a sensor. | Yes | Unknown | Unknown | Yes | In Consideration | Unknown |
| A.5 | Configuration Logging | An interface that provides operations to collect configuration information or (run-time) messages from a sensor (or a sensor network) and output into log files or specified media which can be used by routine troubleshooting and in incident handling. | Yes | Unknown | Unknown | Unknown | In Consideration | Unknown |
| A.6 | Instrument Monitoring | An interface that provides operations to check the state of a sensor or a sensor network which can be done periodically or when triggered by events. | Yes | Unknown | Yes | Yes | In Consideration | Yes |
| A.7 | (Parameter) Visualisation | An interface that provide operations to output the values of parameters and measured variables a display device. | Yes | Unknown | Unknown | Yes | Not Applicable | Unknown |
| A.8 | *(Real-Time) (Parameter/Data) Visualisation* | A specialisation of (Parameter) Visualisation which is subject to a real-time constraint. | Unknown | Unknown | Unknown | Yes | Not Applicable | Unknown |
| A.9 | Process Control | An interface that provide operations to receive input status, apply a set of logic statements or control algorithms, and generate a set of analogy and digital outputs to change the logic states of devices. | Yes | Unknown | Unknown | Yes | Not Applicable | No |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| A.10 | Data Collection | An interface that provides operations to obtain digital values from a sensor instrument, associating consistent timestamps and necessary metadata. | Yes | Yes | Yes | Yes | Not Applicable | Yes |
| A.11 | *(Real-Time) Data Collection* | A specialisation of Data Collection which is subject to a real-time constraint. | Yes | Yes | Unknown | Yes | Not Applicable | Yes |
| A.12 | Data Sampling | An interface that provides operations to select a subset of individuals from within a statistical population to estimate characteristics of the whole population. | No | Unknown | Unknown | Yes | Not Applicable | No |
| A.13 | Noise Reduction | An interface that provides operations to remove noise from scientific data. | Yes | Unknown | Unknown | Yes | Not Applicable | Yes |
| A.14 | Data Transmission | A interface that provides operations to transfer data over communication channel using specified network protocols. | Yes | Yes | Yes | Yes | Not Applicable | Yes |
| A.15 | *(Real-Time) Data Transmission* | A specialisation of Data Transmission which handles data streams using specified real-time transport protocols. | Yes | Yes | Unknown | Yes | Not Applicable | Yes |
| A.16 | Data Transmission Monitoring | An interface that provides operations to check and report the status of data transferring process against specified performance criteria. | Yes | Unknown | No | No | Not Applicable | No |
| **B** | **Data Curation Sub-System** | | | | | | | |
| **No** | **Functions** | **Definitions** | **ICOS** | **EPOS** | **EMSO** | **EISCAT-3D** | **LifeWatch** | **EURO-Argo** |
| B.1 | Data Quality Checking | An interface that provides operations to detect and correct (or remove) corrupt, inconsistent or inaccurate records from data sets. | Yes | Yes | Unknown | Yes | Not Applicable | Yes |
| B.2 | Data Quality Verification | An interface that provides operations to support manual quality checking. | Yes | Unknown | Unknown | Unknown | Not Applicable | Yes |
| B.3 | Data Identification | An interface that provides operations to assign (global) unique identifiers to data contents. | Yes | Yes | Yes | Unknown | Not Applicable | Unknown |
| B.4 | Data Cataloguing | An interface that provides operations to associate a data object with one or more metadata objects which contain data descriptions. | Unknown | Yes | Yes | Unknown | Not Applicable | Unknown |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| B.5 | Data Product Generation | An interface that provides operations to process data against requirement specifications and standardised formats and descriptions. | Yes | Yes | Yes | Yes | Not Applicable | Yes |
| B.6 | Data Versioning | A interface that provides operations to assign a new version to each state change of data, allow to add and update some metadata descriptions for each version, and allow to select, access or delete a version of data. | Yes | Unknown | Unknown | Unknown | Not Applicable | Unknown |
| B.7 | Workflow Enactment | An interface that provide operations or services to interprets predefined process descriptions and control the instantiation of processes and sequencing of activities, adding work items to the work lists and invoking application tools as necessary. | No | Yes | Unknown | Yes | Not Applicable | No |
| B.8 | Data Storage & Preservation | An interface that provides operations to deposit (over long-term) the data and metadata or other supplementary data and methods according to specified policies, and make them accessible on request. | Yes | Yes | Yes | Yes | Not Applicable | Yes |
| B.9 | Data Replication | An interface that provides operation to create, delete and maintain the consistency of copies of a data set on multiple storage devices. | No | Yes | Unknown | Yes | Not Applicable | Yes |
| B.10 | Replica Synchronisation | An interface that provides operations to export a packet of data from on replica, transport it to one or more other replicas and to import and apply the changes in the packet to an existing replica. | No | Unknown | No | Unknown | Not Applicable | Yes |
| **C** | **Data Access Sub-System** | | | | | | | |
| **No** | **Functions** | **Definitions** | **ICOS** | **EPOS** | **EMSO** | **EISCAT-3D** | **LifeWatch** | **EURO-Argo** |
| C.1 | Access Control | An interface that provides operations to approve or disapprove of access requests based on specified access policies. | Unknown | Yes | Unknown | Yes | Unknown | Unknown |
| C.2 | Resources Annotation | An interface that provides operations to create, change or delete a note that reading any form of text, and to associate them with a computational object. | No | No | No | No | Yes | No |
| C.3 | *(Data) Annotation* | A specialisation of Resource Annotation which allows to associate an annotation to a data object. | Yes | Yes | Yes | No | Yes | No |
| C.4 | Metadata Harvesting | An interface that provides operations to (regularly) collect metadata (in agreed formats) from different sources. | Unknown | Unknown | Yes | No | Unknown | No |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| C.5 | Resource Registration | An interface that provides operations to create an entry in a resource registry and insert resource object or a reference to a resource object in specified representations and semantics. | – | – | – | – | – | – |
| C.6 | *(Metadata) Registration* | A specialisation of Resource Registration, which registers a metadata object in a metadata registry. | Unknown | Yes | Yes | No | Unknown | No |
| C.7 | *(Identifier) Registration* | A specialisation of Resource Registration, which registers an identifier object in an identifier registry. | Unknown | Unknown | Yes | No | Unknown | No |
| C.8 | *(Sensor) Registration* | A specialisation of Resource Registration which registers a sensor object to a sensor registry. | Unknown | Unknown | Yes | No | Yes | No |
| C.9 | Data Conversion | An interface that provides operations to convert data from one format to another format. | Yes | Yes | Yes | Yes | Yes | Yes |
| C.10 | Data Compression | An interface that provides operations to encode information using reduced bits by identifying and eliminating statistical redundancy. | No | No | No | No | Yes | No |
| C.11 | Data Publication | An interface that provides operations to provide clean, well-annotated, anonymity-preserving datasets in a suitable format, and by following specified data-publication and sharing  policies to make the datasets publically accessible or to those who agree to certain conditions of use, and to individuals who meet certain professional criteria. | Yes | Unknown | Yes | Unknown | Yes | Yes |
| C.12 | Data Citation | An interface that provides operations to assign an accurate, consistent and standardised reference to a data object, which can be cited in scientific publications. | No | Unknown | Yes | No | Unknown | No |
| C.13 | Semantic Harmonisation | An interface that provides operations to unify similar data (knowledge) models based on the consensus of collaborative domain experts to achieve better data (knowledge) reuse and semantic interoperability. | No | Yes | Yes | No | Yes | No |
| C.14 | Data Discovery and Access | An interface that provides operations to retrieve requested data from a data resource by using suitable search technology. | Yes | Yes | Yes | Yes | Yes | Unknown |
| C.15 | Data Visualisation | An interface that provides operations to display visual representations of data. | Yes | Yes | Yes | Yes | Yes | Yes |

| D | Data Processing Sub-System | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **No** | **Functions** | **Definitions** | **ICOS** | **EPOS** | **EMSO** | **EISCAT-3D** | **LifeWatch** | **EURO-Argo** |
| D.1 | Data Assimilation | An interface that provides operations to combine observational data with output from a numerical model to produce an optimal estimate of the evolving state of the system. | Yes | Unknown | Unknown | Unknown | Unknown | Not Applicable |
| D.2 | Data Analysis | An interface that provides operations to inspect, clean, transform data, and to provide data models with the goal of highlighting useful information, suggesting conclusions, and supporting decision making. | Yes | Yes | Yes | Yes | Yes | Not Applicable |
| D.3 | Data Mining | An interface that provides operations to support the discovery of patterns in large data sets. | Yes | Unknown | No | No | Yes | Not Applicable |
| D.4 | Data Extraction | A interface that provides operations to retrieve data out of (unstructured) data sources, including web pages ,emails, documents, PDFs, scanned text, mainframe reports, and spool files. | Yes | Unknown | Unknown | Yes | Yes | Not Applicable |
| D.5 | Scientific Modelling and Simulation | An interface that provides operations to support of the generation of abstract, conceptual, graphical or mathematical models, and to run an instance of the model. | Yes | Yes | Unknown | Unknown | Yes | Not Applicable |
| D.6 | *(Scientific) Workflow Enactment* | A specialisation of Workflow Enactment, which support of composition and execution a series of computational or data manipulation steps, or a workflow, in a scientific application. Important processes should be recorded for provenance purposes. | No | Unknown | No | No | Yes | Not Applicable |
| D.7 | (Scientific) Visualisation | An interface that provides operations to graphically illustrate scientific data to enable scientists to understand, illustrate and gain insight from their data. | Unknown | Yes | Yes | Yes | Yes | Not Applicable |
| D.8 | Service Naming | An interface that provides operations to encapsulate the implemented name policy for service instances in a service network. | No | Unknown | No | No | Yes | Not Applicable |
| D.9 | Data Processing | An interface that provides operations to initiate the calculation and manage the outputs to be returned to the client. | No | Unknown | No | No | Yes | Not Applicable |
| D.10 | Data Processing Monitoring | An interface that provides operations to check the states of a running service instance. | No | Unknown | No | No | Yes | Not Applicable |

| E | Community Support Sub-System | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **No** | **Functions** | **Definitions** | **ICOS** | **EPOS** | **EMSO** | **EISCAT-3D** | **LifeWatch** | **EURO-Argo** |
| E.1 | Authentication | An interface that provides operations to verify a credential of a user. | Yes | Yes | Unknown | Yes | Yes | Unknown |
| E.2 | Authorisation | An interface that provides operations to specify access rights to resources. | Yes | Yes | Yes | Yes | Yes | Yes |
| E.3 | Accounting | An interface that provides operation to measure the resources a user consumes during access for the purpose of capacity and trend analysis, and cost allocation. | No | Unknown | Yes | No | Unknown | No |
| E.4 | *(User) Registration* | A specialisation of Resource Registration which registers a user to a user registry. | No | Unknown | Unknown | No | Yes | Unknown |
| E.5 | Instant Messaging | An interface that provides operation for quick transmission of text-based messages from sender to receiver. | No | Unknown | No | No | Yes | No |
| E.6 | (Interactive) Visualisation | An interface that provides operations to enable users to control of some aspect of the visual representations of information. | No | Yes | Yes | Yes | Yes | No |
| E.7 | Event Notification | An interface that provide operations to deliver message triggered by predefined events. | No | Yes | Yes | No | Yes | No |

Note, the italicised texts in the table distinguish a specialisation from a particular function.

For the purpose of definition, most of operations are defined as activities performed on an individual object. Sometimes, bulk operations are requested to handle a collection of objects in order to achieve performance. Once requested, bulk functions can be added-on, which will not change the concept of a function fundamentally.

The consistency and completeness of above list of functions will be examined in a follow-on task, T3.3, where we will use ODP activities diagrams to model each process, and to identify missing functional elements. For example, an *Authorisation function* in a ***community support sub-system*** which defines the access policies implies an *Access Control* function in a ***data access sub-system*** to enforce the access policies, thus an *Access Control* function will be added when this is evaluated as necessity.

# 4 COMMON COMMUNITIES

Above analysis helps us gain better understanding about the common architectural characteristics and functionalities by examining the ENV RIs from the aspects of *Engineering* and *Computational Viewpoints*. In this section, we look at the ENV RIs from the ODP *Enterprise Viewpoint*.

The *Enterprise Viewpoint* concerns about the organisational and social context, and scientific processes. It captures the purpose, scope and policies of a system. In order to do that, the system is represented by one or more *enterprise objects* within a **community**, and by the **roles** in which these objects are involved. Using these concepts, in the following, we identify the common *communities* of ENV RIs and community *roles*.

## 4.1 Common Communities

We distinguish 4 communities: ***Data Acquisition***, ***Data Management***, ***Data Service Provision***, and ***Data User***. The division of the 4 communities is based on their main objectives and activities.

- **Data Acquisition Community**, who collects raw data and bring (streams of) measures into a system;
- **Data Management Community**, who curates the scientific data, maintains and archives them, produces various data products with metadata, and makes them public accessible;
- **Data Service Provision Community**, who provides various services, applications and software/tools to link, and recombine data and information in order to derive knowledge;
- **Data User Community**, who make use of the data and service products.

## 4.2 Common Community Roles

Now, we can examine for each *community* which *roles* they may have. Table 4.1 lists common community roles which are either identified from the explicit descriptions in the documentation of ENV RIs or derived from computational functions provided by ENV RIs.

Table 4.1: The Common Roles

| **RA** | **Data Acquisition Community** | | | | | | |
|---|---|---|---|---|---|---|---|
| **No** | **Roles** | **ICOS** | **EPOS** | **EMSO** | **LifeWatch** | **EISCAT-3D** | **EURO-Argo** |
| RA.1 | Ecosystem and environmental resource managers | Unknown | Unknown | Unknown | Yes | Unknown | Unknown |
| RA.2 | Conservation managers | Unknown | Unknown | Unknown | Yes | Yes | Yes |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| RA.3 | Designer for measurements and monitoring models | Yes | Unknown | Yes | No | Yes | No |
| RA.4 | Technician for the development and deployment of the sensor and sensor network | Yes | Unknown | Yes | No | Yes | Yes |
| RA.5 | Technician for the operation and maintenance of the sensor and sensor network | Yes | Yes | Yes | No | Yes | Yes |
| RA.6 | Observer/Measurer/Data collector | Yes | Yes | Yes | No | Yes | Yes |
| RA.7 | Research scientists in data quality control | Yes | Unknown | Unknown | No | Unknown | Yes |
| **RB** | **Data Management Community** | | | | | | |
| **No** | **Roles** | **ICOS** | **EPOS** | **EMSO** | **LifeWatch** | **EISCAT-3D** | **EURO-Argo** |
| RB.1 | Storage Manager | Yes | Yes | Yes | No | Yes | Yes |
| RB.2 | Curator/Data Manager | Yes | Yes | Yes | No | Yes | Yes |
| RB.3 | Data Publisher | Yes | Yes | Yes | Yes | No | No |
| RB.4 | External Data Provider | Yes | Unknown | Yes | No | No | No |
| **RC** | **Data Service Provision Community** | | | | | | |
| **No** | **Roles** | **ICOS** | **EPOS** | **EMSO** | **LifeWatch** | **EISCAT-3D** | **EURO-Argo** |
| RC.1 | Data Provider | Yes | Yes | Yes | Yes | No | No |
| RC.2 | Data Service Provider | Yes | Yes | Yes | Yes | No | Yes |
| RC.3 | Other RIs and Networks with interests overlapping domain | Yes | Yes | Yes | Yes | Yes | Yes |

| RD | Data User Community | | | | | | |
|---|---|---|---|---|---|---|---|
| No | Roles | ICOS | EPOS | EMSO | LifeWatch | EISCAT-3D | EURO-Argo |
| RD.1 | Internal Scientist/researcher who perform in-house experiments/analyses | Yes | Unknown | Yes | Yes | Yes | No |
| RD.2 | External Scientist/researcher | Yes | Yes | Yes | Yes | No | Yes |
| RD.3 | Technologist/engineer | Yes | Yes | Yes | Yes | Yes | No |
| RD.4 | Education/trainee | Yes | Yes | Yes | Yes | No | No |
| RD.5 | Police/decision maker | Yes | Yes | Yes | Yes | Yes | Yes |
| RD.6 | Private sector (Industry investor/consultant) | Yes | Yes | Yes | Yes | No | No |
| RD.7 | General public/media/citizen (scientists) | Yes | Yes | Yes | Yes | Yes | Yes |

Lacking of sufficient resources, the analysis here is very brief. We leave the unfilled spaces for future explorations.

# 5 CONCLUSION

The goal of this investigation was to identify the common requirements of the ENV RIs. Throughout the study, ODP has been used as the analysis framework, which serves as a uniform platform for interpretation and discussion to ensure a unified understanding. From the aspect of the ODP *Engineering Viewpoint*, the architectural characteristics of the RIs have been examined, and 5 common *sub-systems* have been identified: *sub-systems* of **data acquisition**, **curation**, **access**, **processing** and **community support**. Secondly, from the aspect of the ODP *Computational Viewpoint*, we looked at each of the 6 RIs in details and identified the common functions and embedded computations they provided. Matrices has been used for comparison. Definitions of functionalities have been provided. Finally, from the aspect of the ODP *Enterprise Viewpoint*, we have identified 4 common *communities*, and derived the community *roles*.

The results from this study can be delivered as an input to a design or an implementation model. Common services can be provided in the light of the common analysis, which can be widely applicable to various environmental research infrastructures.

There are several elements which could be extended in future work:

- Due to time limitation, only 3 ODP viewpoints have been explored. In future work, analysis from the aspect of the ODP *Information Viewpoint* can be conducted where the requirements for a common information model can be investigated; and from the ODP *Technology Viewpoint*, shared technologies can be identified.
- The analysis from the aspect of ODP *Enterprise viewpoint* has only examined the *communities* and their *roles*. It will be useful to further derive the community *behaviours* (in term of activity processes), and to address community *policy* issues.
- More substantially, how the findings from this study can be better applied to support ENV RIs and other environmental research infrastructures is still an open question.

# 6 ACKNOWLEDGEMENTS

Much help and resources have been obtained from the following people and projects, great thanks to them:

- Esa Turunen, Anders Tjulin and Ingemar Haggstrom from EISCAT-3D project;
- Thierry Carval from Euro-Argo project;
- Jérôme Tarniewicz from ICOS project;
- Robert Huber from PANGAEA project;
- Laura Beranzoli from EMSO project;
- Alex Hardisty, Nicola Fiore and Giuseppe Corriero from LifeWatch project; and
- Malgorzata Krakowian and Gergely Sipos from EGI project.

# 7 REFERENCES

[1] ISO/IEC, "ISO/IEC 10746-1: Information technology--Open Distributed Processing--Reference model: Overview," International Standard, 1998.

[2] ISO/IEC, "ISO/IEC 10746-2: Information technology--Open Distributed Processing--Reference model: Foundations," International Standard, 2009.

[3] ISO/IEC, "ISO/IEC 10746-3: Information technology--Open Distributed Processing--Reference model: Architecture," International Standard, 2009.

[4] ISO/IEC, "ISO/IEC 10746-4: Information technology--Open Distributed Processing--Reference model: Architecture Semantics," International Standard, 1998.

[5] P. F. Linington, Z. Milosevic, A. Tanaka, and A. Vallecillo, Ed., *Building Enterprise Systems with ODP*. CRC Press, 2012.

[6] E. McKay, and I. McCrea, "EISCAT_3D The next generation European Incoherent Scatter radar system: Final Design Study Report Deliverable D11.1," EISCAT-3D Deliverable, 2009.

[7] D. McKay, I. Finch, and I. McCrea, "EISCAT-3D Deliverable D8.3: Data Archiving And Distribution," EISCAT-3D Deliverable, 2008.

[8] IVM, DJM et al, "EISCAT-3D Deliverable 8.1: Stage 2 Report -- WP8 Data segment for EISCAT 3D Radar," EISCAT-3D Deliverable, 2006.

[9] B. Gustavsson, "Visualization for EISCAT 3D," EISCAT-3D Deliverable, 2006.

[10] *Euro-Argo website*: http://www.euro-argo.eu/About-us, Retrieved Dec. 2012.

[11] "Argo Data Management," in *EURO-Argo 1st User workshop*, Presentation, Southampton, 2008.

[12] L. P. de la Villéon, "EURO-Argo: A new European Research Infrastructure," in *ENVRI WP3-WP4 Meeting*, Presentation, Vienna, 2012.

[13] ICOS, "Integrated Carbon Observing System: Stakeholders Handbook 2012," ICOS Design Report, 2012.

[14] J. Tarniewicz, "ICOS: Integrated Carbon Observing System," in *ENVRI Use Case Meeting in Helsinki*, Presentation, Helsinki, 2012.

[15] *EMSO website*: http://www.esonet-noe.org/Gallery/Movies/Deep-sea-observatories-internet-in-the-ocean, Retrieved Dec. 2012.

[16] "Orfeus FDSN Report 2004," Report, 2004.

[17] P. Martin, "EPSO Answer," ENVRI wiki notes, unpublished, 2012.

[18] *Orfeus website: Data Quality*: http://www.orfeus-eu.org/Data-info/dataquality.html, Retrieved Dec. 2012.

[19] P. Martin, "EPSO Use Case," ENVRI wiki notes, unpublished, 2012.

[20] K. Jeffery, T. L. Hoffmann, "Report on EPOS e-Infrastructure Requirements," Report, 2011.

[21] M. B. de Bianchi, and J. Saul, "EIDA: European Integrated Data Archives," Presentation, 2012.

[22] D. Bailo, and T. L. Hoffmann, "D6.2 Annex 2: EPOS use cases," Report, 2011.

[23] *Orfeus website: Software*: http://www.orfeus-eu.org/Software/softwarelib.html, Retrieved 2012.

[24] R. Sleeman, T. van den Hazel, A. Spinuso, and L. Trani, "Integrating seafloor and land-based seismic waveform data at ORFEUS Data Center," in *ORFEUS Observatory Coordination Workshop, 25-28 May 2011*, Presentation, 2011.

[25] V. Hernandez-Ernst, et al., "LIFEWATCH. Deliverable 5.1.3: Data & Modelling Tool Structures -- Reference Model," LIFEWATCH Deliverable, 2010.